AGI logic tutorial

YKY

December 25, 2020

Summary

AGI

Contents

1 Background

optimization problem,

$$\text{maximize: } \int_0^\infty R\, dt \tag{1}$$

where $R(t) =$ reward at time t. $\int_0^\infty$ **time horizon**.

N (1)

Architecturally, the AI is a **dynamical system** that constantly updates its "state" $\boldsymbol{x}$ via: [*]

$$\dot{\boldsymbol{x}} = \boldsymbol{f}(\boldsymbol{x}) \tag{2}$$

$$\boldsymbol{x}_{t+1} = \boldsymbol{F}(\boldsymbol{x}_t) \tag{3}$$

$\boldsymbol{F}$ transition function.

$$\overset{\boldsymbol{F}}{\circlearrowleft}\ \boldsymbol{x} \tag{4}$$

Our goal is to **learn** the function $\boldsymbol{F}$, implemented as a **deep neural network**. $\boldsymbol{F}$

1.1

$\neq$

(5)

architecture (4) Richard Sutton

†

Wright brothers "wing warping" "plane" 100 dominant design.

---

[*] Part of the state $\boldsymbol{x}$ contains **sensory input** and **action output** that allow the AI to interact with the external environment.

[†]AGI episodic memory minimalist architecture.

$\boldsymbol{F}$ inductive bias AGI Sutton
bias

Richard Sutton (1949-) 

2 Structure of logic

The central tenet of my theory is that the state $\boldsymbol{x}$ of the AI system is consisted of **logic propositions** and that $\boldsymbol{F}$ plays the role of the **logic consequence** operator $\vdash$:

$$\boxed{\text{propositions}} \mathrel{\vdash\!\!\xrightarrow{\boldsymbol{F}}} \boxed{\text{propositions}} \tag{6}$$

So our goal now is to elucidate the structure of $\vdash$. Currently the most elegant formulation is given by **categorical logic** or **topos theory**.

"synthesize" ideas ideas

AGI Curry-Howard isomorphism....

3 Curry-Howard correspondence

Curry-Howard isomorphism

(syntax) $A \Rightarrow B$

$$\begin{array}{cc} \boxed{\text{logic}} & A \Longrightarrow B \\ \hdashline \boxed{\text{program}} & \blacksquare \overset{f}{\longmapsto} \blacksquare \end{array} \tag{7}$$

(proof) A map B "proof witness"

$\blacksquare$.

$f(x) = x + 2$

$$\frac{f : \mathbb{R} \longrightarrow \mathbb{R}}{x \longmapsto x + 2} \tag{8}$$

$$x \quad \Longrightarrow x + 2 \tag{9}$$

x (witness)

$$\frac{\boxed{x}}{\blacksquare} \tag{10}$$

This is called the **Brouwer-Heyting-Kolmogorov (BHT) interpretation**. Brouwer-Heyting-Kolmogorov-Schönfinkel-Curry-Meredith-Kleene-Feys-Gödel-Läuchli-Kreisel-Tait-Lawvere-
Howard-de Bruijn-Scott-Martin-Löf-Girard-Reynolds-Stenlund-Constable-
Coquand-Huet-Lambek ....

Curry-Howard isomorphism 1990s Lambek **category theory** Curry-Howard- Lambek. **- -**

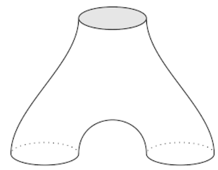
John Baez (1961-)

Curry-Howard isomorphism states A transitions $\overset{f}{\mapsto}$

logic	**computation**	**category theory**	**physics**	**topology**
proposition	type	object	system	manifold
proof	term	morphism	process	cobordism

(11)

Curry-Howard Lambek John Baez & M. Stay *Physics, Topology, Logic and Computation: a Rosetta stone* [2010]. physics Hilbert space operators topology cobordism “pair of pants”



(12)

In string theory strings string

3.1 Type theory

program **computation** type theory.

```
define length(s: String): Integer = { .... }
```
(13)

length() String Integer.

$$f : A \to B \tag{14}$$

type theory

$$\overbrace{t}^{\text{term}} : \overbrace{T}^{\text{type}} \tag{15}$$

notation $t : T$ $t \in T$

types terms

type theory type **context**

$$\overbrace{x : A}^{\text{context}} \vdash \overbrace{f(x) : B}^{\text{type assignment}} \tag{16}$$

program “declare” program

$\vdash$ **type assignment** type theory

λ-calculus

program λ-calculus

λ-calculus

$$f(x) \triangleq x^2 \tag{17}$$

λ-

$$f \triangleq \lambda x.\, x^2 \tag{18}$$

λ- f

λ-calculus Alonso Church **substitution** Substitute

Church λ-calculus **Turing machines** AI John McCarthy λ-calculus **Lisp** functional programming language

Curry-Howard correspondence

Curry-Howard type A A type A A (proof, or proof witness)

$A \Rightarrow B$ the function type $A \to B$ B^A type $f : A \to B$. type $A \to B$ (inhabited) $A \Rightarrow B$

3.2 Intuitionistic logic

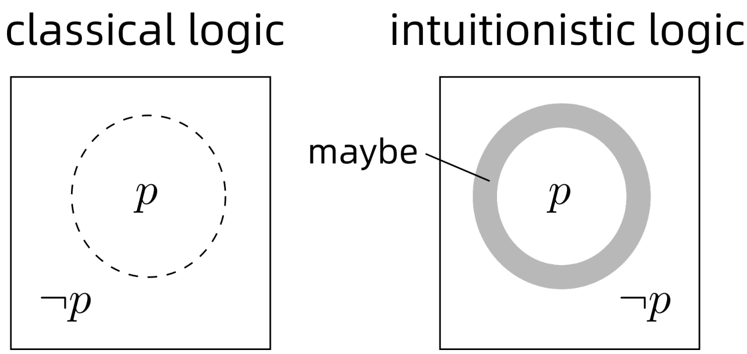
Curry-Howard isomorphism type theory **intuitionistic logic** (law of excluded middle, LEM) **double negation** $\neg\neg p \Rightarrow p$.

$p \vee \neg p$ $p \vee \neg p$ p $\neg p$

axiom of choice $\Rightarrow$ law of excluded middle. axiom of choice

Topological interpretation

open sets Hausdorff p $\overline{p}$ open $\neg p$ p **interior** $\neg p \triangleq \overline{p}^{\circ}$. $p \cup \neg p \neq$ Universe: *



(19)

3.3 Higher-order logic

Propositional logic

p, q $p \wedge q, p \vee q, p \Rightarrow q, \neg p$.

First-order logic

$$\overbrace{\text{IsHuman}}^{\text{predicate}}(\overbrace{\text{John}}^{\text{object}}). \tag{20}$$

Predicate objects predicates

$$\text{Loves}(\text{John}, \text{Mary}). \tag{21}$$

First-order $\forall, \exists$ objects Mary

$$\forall x.\ \text{Loves}(x, \text{Mary}) \tag{22}$$

first-order logic predicates second-order logic.

$$\forall p.\ p(\text{Good General}) \Rightarrow p(\text{Napoleon}). \tag{23}$$

p predicates

* Diagram from the book: *Classical and Non-classical Logics – an introduction to the mathematics of propositions* [Eric Schechter 2005], p.126.

3.4 logic with type theory

Type theory Russell **impredicative**. type theory

Curry-Howard isomorphism type theory $p, q, p \wedge q$ terms Curry-Howard = types = terms.

type theory (first- or higher-order) predicate logic

$$\text{IsHuman}(\text{John}) \tag{24}$$

IsHuman term (truth value) $\in \Omega = \{\top, \bot\}$. IsHuman $\text{Obj} \rightarrow \Omega$ term.

This approach leaves no room to accommodate Curry-Howard isomorphism. To do the latter, we would need Martin-Löf type theory....

3.5 Martin-Löf type theory

Curry-Howard, $A \Rightarrow B$ type:

$$\underset{\Omega}{\overbrace{\text{Human (Socrates)}}^{A}} \Rightarrow \underset{\Omega}{\overbrace{\text{Mortal (Socrates)}}^{B}} \tag{25}$$

Human() Mortal() predicates type theory types. $\rightarrow$ $\Rightarrow$ Curry-Howard type theory

"simple" type theory

- sum type $A + B$
- product type $A \times B$
- function type $A \rightarrow B$

$\vee, \wedge, \Rightarrow$. type theory

Human(Socrates) Human() Socrates arrow $\rightarrow$ arrow

Martin-Löf **type constructors**

dependent sum type Σ

dependent product type Π

Dependent sum $\sum\limits_A B$ B depends on A. family of A $+$ product $A \times B$.

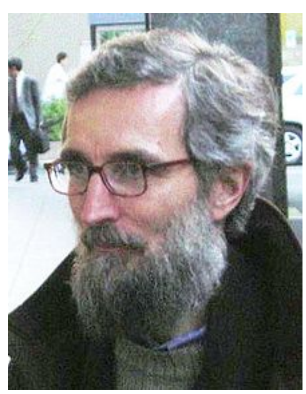
Dependent product $\prod\limits_A B$ B depends on A. family of A $\times$ exponentiation B^A.

Dependent products can be used to define **predicates** such as Human() and Mortal(). They are of type $\mathrm{Obj} \to \Omega = \Omega^{\mathrm{Obj}} = \prod\limits_{\mathrm{Obj}} \Omega$. *

$\sum_A B$ $\prod_A B$ types inhabited $\exists A.B(A)$ $\forall A.B(A)$. $A \times B$ inhabited $B(A)$ B^A inhabited A send to B.

Per Martin-Löf (1942-) was the first logician to see the full importance of the connection between intuitionistic logic and type theory.

Per Martin-Löf (1942-)



* Note that "objects" here mean logic objects, not objects in category theory.

3.6 Arithmetic-logic correspondence

$\wedge, \vee$ $\times, +$ fuzzy logic $\min, \max.$ George Boole

$A \Rightarrow B$ B^A *

A	B	$A \Rightarrow B$	B^A
0	0	1	$0^0 = 1$
0	1	1	$1^0 = 1$
1	0	0	$0^1 = 0$
1	1	1	$1^1 = 1$

(26)

Curry-Howard correspondence $\Rightarrow$ "functional interpretation of logical deduction."

table (26) A B truth values types **inhabitants**. Type A inhabitant $\blacksquare$ $\varnothing$. A $= A$ type **cardinality**; The truth valuation of $A = |A|$. $A \Rightarrow B$ $|B^A|$, $\{\blacksquare\}$ $\varnothing$ $\{\blacksquare\}$ $\varnothing$ map, map $\varnothing \mapsto \{\blacksquare\}$

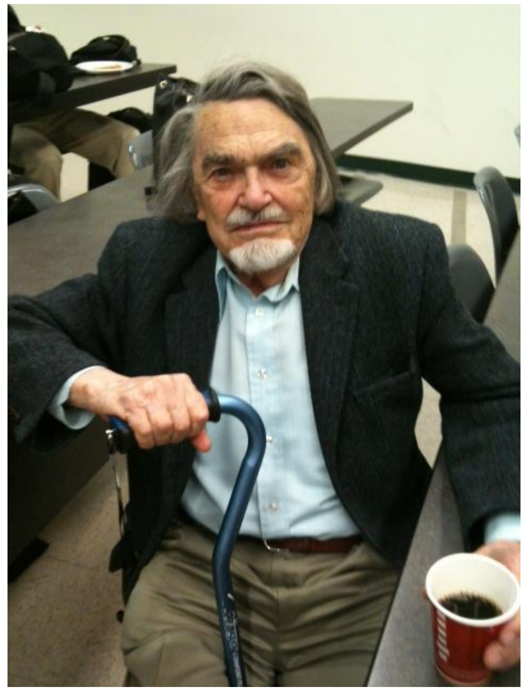
fuzzy logic (§7.1) strict implication $A \multimap B$ (§6.5).

---

* 0^0 1.

4 Topos theory

Joachim Lambek (1922-2014)



type theory category theory Lambek Curry-Howard-Lambek

$$
\begin{tikzcd}
\boxed{\text{logic}} \arrow[rr, dash] \arrow[dr, dash] & & \boxed{\text{programming}} \arrow[dl, dash] \\
& \boxed{\text{category theory}} &
\end{tikzcd}
\tag{27}
$$

Topos **sub-objects**.

topos $\mathcal{C}$ sub-object classifer Ω $X \to \Omega \cong$ sub-objects of X. X $X \to \Omega$ **represent**.

Set topos Ω $\{\top, \bot\}$. $X \to \Omega$ X $X \xrightarrow{\text{mathematician}} \Omega$

Topos theory commutative diagram

$$
\begin{tikzcd}
X \arrow[r, "!"] \arrow[d, tail, "m"'] & 1 \arrow[d, "\text{true}"] \\
Y \arrow[r, "\chi_m"'] & \Omega
\end{tikzcd}
\tag{28}
$$

$X \xrightarrow{!} 1$ unique arrow X 1. 1 terminal object

$1 \xrightarrow{\text{true}} \Omega$ $\top$ $\bot$ $\top$, "true" arrow.

$X \overset{m}{\rightarrowtail} Y$ **monic** arrow **Set** **inclusion** map $X \overset{m}{\hookrightarrow} Y$. X Y $X \subseteq Y$.

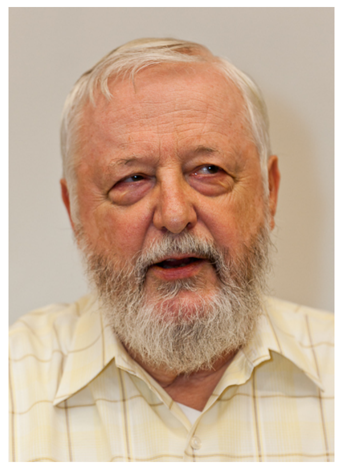
$Y \xrightarrow{\chi_m} \Omega$ **characteristic function**, $e \in X \subseteq Y$ $\chi(e)$ 1 0. χ_m commute. χ_m $\ulcorner m \urcorner$.

Conceptual Mathematics Lawvere topos

Set diagram topos **generalize** **Set**

Topos category $P(x), \forall x, \exists x,$ Lawvere

William Lawvere (1937-)



(28):

$$\begin{tikzcd}
X \arrow[r, "!"] \arrow[d, rightarrowtail, "m"'] \arrow[dr, phantom, "\lrcorner", very near start] & 1 \arrow[d, "\text{true}"] \\
Y \arrow[r, "\ulcorner m \urcorner"'] & \Omega
\end{tikzcd} \tag{29}$$

$\begin{smallmatrix}1\\ \downarrow\\ \Omega\end{smallmatrix}$ **generic subobject**. **pull back** square $\lrcorner$ $\begin{smallmatrix}X\\ \downarrow\\ Y\end{smallmatrix}$ sub-object. We say that the **property** of being a sub-object is **stable under pullbacks**.

4.1 $\wedge$ and $\Rightarrow$ in a topos

Material implication compare truth values, fuzzy

Exponentiation B^A

4.2 $\forall$ and $\exists$ as adjunctions

Let Forms($\vec{x}$) denote the set of formulas with only the variables $\vec{x}$ free.

Then there is a trivial operation of adding an additional dummy variable y:

$$* : \text{Forms}(\vec{x}) \to \text{Forms}(\vec{x}, y) \tag{30}$$

taking each formula $\phi(\vec{x})$ to itself.

It turns out that $\exists$ and $\forall$ are adjoints to the map $*$:

$$\exists \dashv * \dashv \forall \tag{31}$$

4.3 Classifying topos $\leftrightharpoons$ internal language

objects morphisms

Lambek

types $\leftrightsquigarrow$ objects

terms $\leftrightsquigarrow$ morphisms

We have the following transformations betweeen two formalisms:

$$\boxed{\text{topos}}\ \mathcal{C} \underset{\text{classifying topos}}{\overset{\text{internal language}}{\rightleftarrows}} T\ \boxed{\text{type theory}}\ . \tag{32}$$

In other words,

$$\mathcal{C} = \mathcal{C}\ell(T), \quad T = \text{Th}(\mathcal{C}). \tag{33}$$

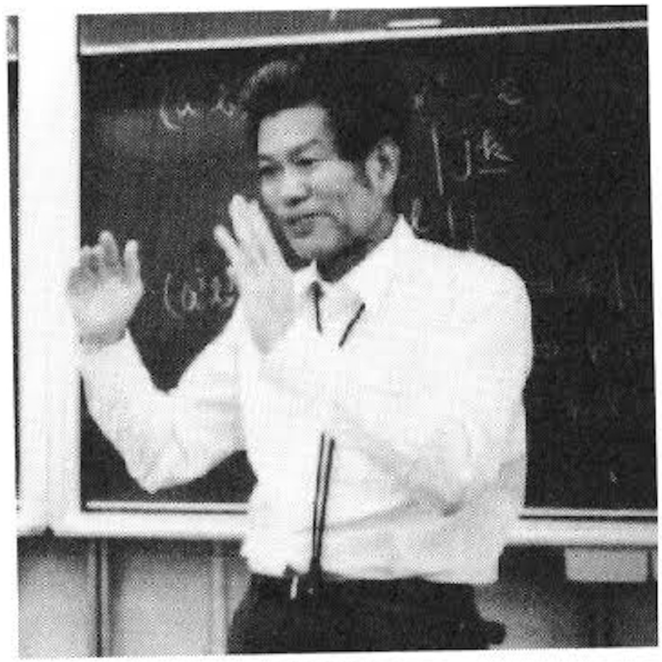
4.4 Sheaves and topos

Some **Set**-valued functors are **representable**, ie, isomorphic to a hom-functor.

Functors $\mathcal{C} \to$ **Set** are called **pre-sheaves** on $\mathcal{C}$.

Sheaves capture "indexing".

4.5 Yoneda lemma



Nobuo Yoneda (1930-1996)

$\mathcal{C}$ A morphism, $A \to \bullet$. A

Set $1 \to X$ 1

$\mathbb{R} \to X$ X $\mathbb{R}$

ordered set $(\mathbb{R}, \leq)$ $0 \to x$ x **positive**.

$\bullet \to A$ A

Set $X \to 2$ 2 X , $\wp(X)$.

Top 2 open set closed set $X \to 2$ X Opens(X).

How sheaves gives rise to representables.

4.6 Model theory, functorial semantics

$$a \cdot b \longmapsto \llbracket a \rrbracket \, ; \, \llbracket b \rrbracket \tag{34}$$

4.7 Generalized elements and forcing

ϕ $A \xrightarrow{\phi} \Omega$ $\Omega = \{\top, \bot\}$.

$\phi(x)$ x A In category theory, we use the terminal object 1 to "pick out" elements of A, as follows:

$$1 \xrightarrow{x} A \xrightarrow{\phi} \Omega. \tag{35}$$

In **Set**, 1 $x : 1 \to A$ A

C 1

$$C \xrightarrow{x} A \xrightarrow{\phi} \Omega. \tag{36}$$

$x : C \to A$ A **generalized element**.

C **forces** $\phi(x)$, notation $C \Vdash \phi(x)$.

$\phi(x)$ is true **at stage** C possible-world semantics .

4.8 Kripke-Joyal semantics

An "internal" way to interpret type theory in a topos is where a formula ϕ in context $x_1 : A_1, ..., x_n : A_n$ is interpreted as a subobject of $A_1 \times ... \times A_n$. This has the disadvantage that the most pleasant illusion of "elements" is totally lost.

Kripke-Joyal semantics generalized elements

Generalized element $I \xrightarrow{a} A \xrightarrow{\phi} \Omega$, $a \Vdash \phi$.

4.9 Cohen's (dis)proof of Continuum Hypothesis

AGI

Continuum hypothesis (CH):

$$2^{\aleph_0} = \aleph_1 \tag{37}$$

$[0, 1]$ $2^{\aleph_0}$

1878 Cantor CH

1900 Hilbert 23

Hilbert bug

1938 Gödel ZF + CH is consistent ZF cannot disprove CH

1963 Paul Cohen ZF cannot prove CH

"forcing"

Paul Cohen (1934-2007) 

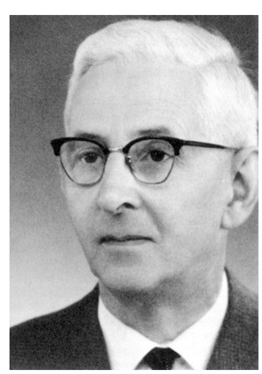
4.10 Kleene realizability

5 Intuitionistic logic

In 1933, Gödel proposed an interpretation of intuitionistic logic using possible-world semantics.

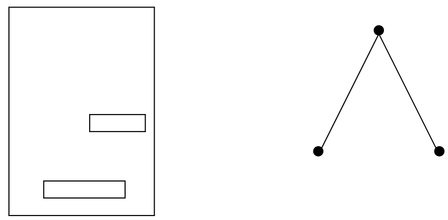
In topos theory $A \Rightarrow B$ is adjoint (via the hom-product adjunction) to $A \vdash B$, which is "okay" because it is independent of which implication (material or strict) we are using.

5.1 Heyting algebra

Arend Heyting (1898-1980)  (38)

1930 Heyting constructive mathematics axiomatization, called **intuitionistic logic** (IL). Heyting algebra IL Boolean algebra

(lattice) Heyting algebra



(39)

Stone duality (every Boolean algebra is isomorphic to a topology of open sets) **Priestley**

$\longrightarrow$ Heyting implication $\longrightarrow$

In a topos $\mathbb{E}$, the subobject $\mathrm{Sub}_{\mathbb{E}}(A)$ is a **poset** that admits **Heyting implication**. The Heyting implication $a \Rightarrow b$ exists for all elements a, b, x such that:

$$x \leq (a \Rightarrow b) \quad \text{iff} \quad (x \wedge a) \leq b. \tag{40}$$

Every Boolean algebra can be a Heyting algebra with the material implication defined as usual: $a \Rightarrow b \equiv \neg a \vee b$.

Heyting algebra is to intuitionistic logic what Boolean algebra is to classical logic. But this may not jibe with the idea of "strict implication".

Under Kripke semantics, the Heyting arrow $\rightarrow$ can be defined by:

$$k \Vdash A \rightarrow B \quad \Leftrightarrow \quad \forall \ell \geq k \; (\ell \Vdash A \Rightarrow \ell \Vdash B) \tag{41}$$

Whereas the "fish-hook" **strict implication** can be defined as "A implies B necessarily":

$$A \strictif B \quad \equiv \quad \Box(A \Rightarrow B) \tag{42}$$

The two can be regarded as equivalent via:

$$\begin{aligned} k \Vdash \Box(A \Rightarrow B) \quad &\Leftrightarrow \quad \forall \ell \geq k \; (\ell \Vdash (A \Rightarrow B)) \\ &\Leftrightarrow \quad \forall \ell \geq k \; (\ell \Vdash A \Rightarrow \ell \Vdash B) \end{aligned} \tag{43}$$

6 Modal logic

Modalities are often conceived in terms of variation over some collection or **possible worlds**.

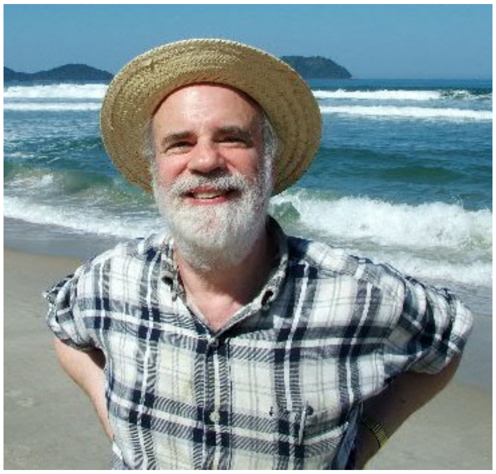
A modal operator (such as $\Box$) in the category **Sheaf**(X) is a **sheaf morphism** $\Box : \Omega \to \Omega$ satisfying 3 conditions, $\forall U \subseteq X$ and $p, q \in \Omega(U)$:

$$\begin{array}{ll} \text{a)} & p \leq \Box(p) \\ \text{b)} & (\Box;\Box)(p) \leq \Box(p) \\ \text{c)} & \Box(p \wedge q) = \Box(p) \wedge \Box(q) \end{array} \tag{44}$$

6.1 Possible-world semantics

Possible-world semantics is also called **intensional semantics**, as opposed to **extensional semantics** where truth values are directly assigned to propositions. Does this idea jibe with the other definition of "intension", ie, as opposed to Leibniz extensionality and also related to intensional logic?

Saul Kripke (1940-)



6.2 Computer implementation of possible worlds

modal logic frame $F = \langle W, R, D, H \rangle$,

W = set of possible worlds = $\{w_1, w_2, ...\}$

$R =$ a relation between worlds, $w_i R w_j$

$D =$ domain of first-order objects

$H : W \to \wp(D)$, for each world specify a subset of objects

To interpret formulas with $\Box$:

$$M \Vdash \Box A\ [w] \quad \Leftrightarrow \quad \forall w' \succeq w.\ M \Vdash A\ [w'] \tag{45}$$

quantify over all w's.

inference data

w p $\forall w.p[w]$.

summary working memory quantify.

6.3 Intensional vs extensional

"Beethoven's 9th symphony" and "Beethoven's choral symphony" has the same **extension** but different **intensions**.

6.4 Intensional logic

Possible-world semantics is also called **intensional semantics**, as opposed to **extensional semantics** where truth values are directly assigned to propositions.

Logic terms differ in intension if and only if it is **possible** for them to differ in extension. Thus, **intensional logic** interpret its terms using possible-world semantics.

6.5 Strict implication

The problem of "material implication"

Material implication $A \Rightarrow B$ $\neg A \vee B$

A	B	$A \Rightarrow B$
0	0	1
0	1	1
1	0	0
1	1	1

(46)

Material implication

$$\Rightarrow \tag{47}$$

---

truth table material implication strict implication

For strict implication to make sense, it is always necessary to invoke possible-world semantics. A strict implication is always **learned** from numerous examples from experience, in accord with the philosophical tradition of "empiricism".

Strict implication is equivalent to material implication over multiple instances. The truth table of material implication agrees with the functional interpretation of implication.

7 Fuzzy logic

Lotfi Zadeh (1921-2017) 

Iranian-Jewish

implication fuzzy implication material implication in Boolean algebra.

fuzzy truth value

set fuzzy proposition

fuzzy value

John

fuzzy value

John

John

imply

$\subseteq$

"Marilyn Monroe is sexy" Marilyn imply sexy subset sexy Marilyn

Sexy(marilyn), Human(john), vs Human(Mathematicians).

What kind of mapping does this require?

7.1 Fuzzy implication

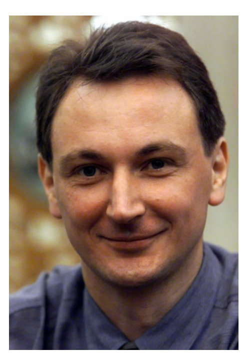
Implication generalize fuzzy logic

7.2 Fuzzy functions?

What are fuzzy functions?

8 Homotopy type theory (HoTT)

Vladimir Voevodsky (1966-2017)



HoTT types (spaces) topological homotopy homotopy type A id_A homotopy **path**.

8.1 HoT levels

...	...
2	2-groupoids
1	groupoids
0	sets
-1	(mere) propositions
-2	contractable spaces

(48)

8.2 What is homotopy?

8.3 Univalence axiom

HoTT set "=" predicate, fuzzy predicate.

univalence axiom = fuzzy truth value binary.

References

Questions, comments welcome ☺